데이터의 지역성을 이용한 빈발구간 항목집합 생성방법

박원환[†] · 박두순[††]

요 약

최근에 대용량의 데이터베이스로부터 연관규칙을 발견하고자 하는 연구가 활발하며, 수량항목에도 적용할 수 있도록 이들 방법을 확장하는 연구도 소개되고 있다. 본 논문에서는 수량항목을 이진 항목으로 변환하기 위하여 빈발구간 항목집합을 생성할 때, 수량 항목의 정의 영역 내에서 특정 영역에 집중하여 발생하는 특성인 지역성을 이용하는 방법을 제안한다. 이 방법은 기존의 방법보다 많은 수의 세밀한 빈발구간 항목들을 생성할 수 있을 뿐만 아니라 세밀도를 판단하여 활용할 수 있는 생성순서 정보도 포함하고 있어, 원 데이터가 가지고 있는 특성의 손실을 최소화할 수 있는 특징이 있다. 인구센서스 등 실 데이터를 사용한 성능평가를 통하여 기존의 방법보다 우수함을 보였다.

A Method for Generating Large-Interval Itemset using Locality of Data

Won-Hwan Park[†] and Doo-Soon Park[††]

ABSTRACT

Recently, there is growing attention on the researches of inducing association rules from large volume of database. One of them is the method that can be applied to quantitative attribute data. This paper presents a new method for generating large-interval itemsets, which uses locality for partitioning the range of data. This method can minimize the loss of data-inherent characteristics by generating denser large-interval items than other methods. Performance evaluation results show that our new approach is more efficient than previously proposed techniques.

1. 서 론

데이터베이스 기술의 발전으로 이를 사용하는 업무가 급속히 늘어나면서 저장되는 데이터 양 또한 폭발적으로 증가하여 대용량화되고 있다. 이러한 대용량화된 데이터베이스에는 사용자가 미처 파악하지 못하는 중요한 정보 또는 지식이 포함되어 있을 수 있으나, 기본적으로 데이터베이스는 일정한 형태의 질의에 빠른 응답을 위한 시스템이므로 현실세계(real world)에서 나타나는 다양한 규칙성(regularity)을 발견하기에는 한계가 있다. 이러한 대용량 데이터베이스에 내재된 유용한 지식을 탐사하는 기술을 데이터마이닝(data mining)이라 하며, 이에 대한 연구가 최근에 활발하게 진행되고 있다.

데이터마이닝은 대량의 실제 데이터, 즉 트랜잭션(transaction)을 발생시킨 특성을 효과적으로 반영하여 의사결정(decision making)에 유용한 정보를 제공한다. 예를 들어 대형 할인매장에서 "노트"와 "연필"을 구매한 고객의 85%가 "지우개"를 같이 구매한다는 연관규칙을 판매 데이터베이스의 자료에서 찾을 수 있다. 이와 같이 데이터마이닝은 상식적인 감(feeling)을 사실(fact)로 전환시킬 수도 있지만, 슈퍼

본 연구는 정보통신부의 ITRC 사업에 의해 수행된 것임
† 정회원, 순천향대학교 대학원 박사과정
†† 정회원, 순천향대학교 정보기술공학부 교수

마켓에서 "아기 기저귀"를 구매한 사람이 "맥주"도 함께 구매한다와 같이 전혀 예상하지 못한 규칙도 탐사할 수 있다. 이처럼 데이터마이닝은 대용량의 데이터베이스 자체가 가지고 있는 모든 규칙을 탐사할 수 있다.

데이터마이닝은 크게 예측(prediction)과 지식탐사(knowledge discovery)로 구분[1]한다. 예측은 특별한 목표에 관심을 가지고 과거의 기록을 기초로 미래의 새로운 사건(case)에 적용하기 위하여 이용하며, 분류(classification), 시계열(time series) 등의 방법이 있다. 반면에 지식탐사는 예측보다 적은 정보로도 가능하며 의사결정지원(decision support)에 적합한 방법이다. 군집화(clustering), 연관규칙(association rules) 등이 해당된다. 지식탐사의 한 분야인 연관규칙(association rules) 탐사에 관한 문제는 Agrawal[2]이 처음 제안한 이후 수많은 연구가 있었고, 최근에도 이 분야의 연구는 활발하게 진행되고 있다.

현실세계에서 발생하는 트랜잭션(또는 사건)은 그 특성을 나타내는 항목(item)들의 집합으로 항목집합(itemset)의 단위로 데이터베이스에 기록되어 대용량화된다. 여기에 자주 발생하는항목집합들 간의 상호관련성을 발견하는 작업을 연관규칙(association rule)이라 한다. 이 분야는 이진 연관규칙(binary association rule), 즉 "한 트랜잭션 내에서 특정 항목이 나타나면 반드시 다른 항목도 그 트랜잭션 내에 함께 나타난다." 라는 형태의 규칙성 발견에 대한 연구가 활발하다[3-7].

센서스 데이터와 같이 수량 속성이 강한 자료의 경우에도 이진 연관규칙 탐사가 가능하다. 그러나 연령, 자녀 수 등 센서스 데이터에 포함된 많은 수량 속성을 무시하고 연관규칙을 탐사한다는 것은 발견할 수 있는 규칙이 극히 제한적이거나 불가능하다. 예들 들어 사람의 연령은 수량 데이터로써 발생영역은 1세에서 120세인데, 이를 하나의 항목, 즉 '연령'을 취하여 연관규칙을 탐사할 경우 발견할 수 있는 규칙은 극히 제한적이다. 반대로 수량항목의 정의영역 내의 각 수치를 그대로 항목으로 설정하여 빈발항목집합을 생성한다면 탐사공간이 너무 넓고, 발생하는 항목이 분산되어 있어 탐사가 불가능하다.

이와 같은 문제는 수량 항목의 정의영역을 여러 개의 소구간으로 분할하여 최소지지도를 만족할 때까지 병합하여 빈발구간 항목집합[1])을 생성[8,9]한다. 빈발구간 항목은 최소지지도를 만족해야 하는데, 이의 판단기준은 해당 구간 내에 데이터의 발생빈도(frequency)에 따라 결정된다. 일반적으로 실세계 데이터들의 발생빈도는 특정 영역에 치우치는 경향, 즉 지역성(locality)를 갖는다.

본 논문에서는 센서스 데이터와 같이 수량적 속성 데이터의 항목을 여러 개의 구간으로 분할할 때 데이터 발생의 지역성을 고려하여 빈발구간 항목집합을 생성하는 방법을 제안한다. 이 방법은 밀도 높은 구간을 중심으로 빈발구간 항목을 생성하므로 원래의 데이터가 가진 특성의 손실을 최소화할 수 있는 특징이 있다.

본 논문의 구성은 제2장에서 연관규칙탐사의 기본 정의, 탐사 알고리즘 그리고 수량 항목에 대한 빈발구간 항목집합 생성방법에 대하여 알아본다. 3장에서는 빈발구간 항목집합을 생성하는 새로운 방법을 제안하고, 그 예를 보인다. 4장에서는 성능을 평가하고, 마지막으로 결론 및 향후과제를 밝힌다.

2. 연관 규칙 탐사

2.1 연관 규칙 정의

2.1.1 빈발 항목집합(large itemset)의 정의

빈발항목집합을 정의하기 위하여 항목집합(I), 트랜잭션(T), 데이터베이스(D)를 다음과 같이 정의한다.

- $I = \{i_1, i_2, i_3, \cdots, i_m\}$, i_j $(j=1, \ldots, m)$는 항목(item)이다.
- T : I의 부분집합($T \subseteq I$)이며, 항목의 중복은 불허한다.
- D : n개의 트랜잭션들 집합이며, 각 트랜잭션은 고유한 번호(Tid)를 가진다.

여기서, 트랜잭션과 다른 모든 항목집합들 내에 있는 항목들은 정렬된 것으로 가정한다.만일 트랜잭션 T가 X의 모든 항목들을 포함한다면($X \subseteq T$), T가 집합 X를 지지한다(support)고 한다. 물론 X는 I의 부분집합이다. X의 지지도를 supp(X)로 표기하며, 이는 X를 지지하는 D에 있는 모든 트랜잭션들의 수

1) 수량항목의 빈발항목집합(large itemset)을 이진항목의 경우와 구분하기 위하여 빈발구간 항목집합(large-interval itemset)이라 정의한다.

를 의미한다.

만일 최소지지도(S_{min})에 대하여 supp(X) ≥ S_{min} 이라면, "*집합 X는 빈발하다*" 라고 하며, 항목들의 집합 X를 빈발항목집합(large itemset 또는 frequent itemset)이라 한다. 최소지지도는 주어지며, D에 대하여 관심 있는 항목만을 고려대상으로 하기 위하여 사용한다. 또한 X가 k개의 항목으로 구성되어 있다면, k-항목집합(k-itemset)라 한다.

2.1.2 연관 규칙(association rule)의 정의

한 트랜잭션에서 발생한 항목들의 양은 고려하지 않으며, 그 항목들의 발생 여부만을 고려하기로 가정한다. 연관규칙(association rule) R을 다음과 같이 정의한다.

R : X → Y

이 때 X와 Y는 서로 같은 원소를 갖지 않는 항목이다. 즉 X, Y⊆I 이고, X∩Y = ϕ이다.

만약 데이터베이스 D의 트랜잭션들 중에서 s %가 X∪Y 를 포함한다면 연관규칙 R : X → Y는 트랜잭션들의 집합 D에서 지지도(support degree) s 를 가지고 있다. 또한만약 D에 있는 트랜잭션들 중에서 c %가 X를 포함하고 있고, 이들이 또한 Y를 포함하고 있다면 연관규칙 R : X → Y는 집합 D에서 신뢰도(confidence degree) c 를 가진다.

2.1.3 연관규칙 탐사의 기본적인 접근방법

연관규칙 탐사 문제와 관련된 알고리즘은 다양하게 연구되었고, 또한 연구되고 있다. 그러나 서로 다른 알고리즘에도 불구하고 이들의 기본적인 스키마는 유사하다. 즉, 연관 규칙을 탐사하기 위하여 데이터베이스에 있는 모든 항목들의 지지도를 계산하여 빈발 항목집합(large itemsets)을 찾고, 이로부터 주어진 신뢰도를 바탕으로 실제의 규칙을 탐사하는 과정으로 이루어지는 2단계 구조이다.

단계 1 : 빈발 항목집합들(large itemsets)을 찾는 단계로 주어진 최소지지도(S_{min}) 이상의 트랜잭션 지지도를 가지는 항목집합들인 빈발항목집합을 찾는 과정

단계 2 : [단계1]에서 생성된 빈발 항목집합을 사용하여 연관규칙을 생성하는 단계

연관 규칙 탐사의 전체성능은 첫 번째 단계에서 결정된다. 데이터베이스 속에 고려대상 빈발 항목의 수는 모든 항목들의 멱집합(power set)의 크기와 같다. 즉 항목들의 수 증가에 대하여 고려해야 할 항목의 크기는 기하급수적으로 증가하여 상당량의 처리시간과 메모리를 요구한다. Apriori[3], AprioriTID [3,10], AprioriHybrid[3], DHP[4], Partition[5], DIC [11], Direct Sampling[5], Sampling Approach[6]등 연관규칙 탐사 알고리즘들 대부분이 이러한 문제의 해결에 중점을 두고 있다.

2.2 연관규칙 탐사 알고리즘

연관규칙 탐사의 대표적인 알고리즘으로 알려진 Apriori는 해당 항목의 발생 유무만 고려하는 이진항목의 탐사에 적합한 알고리즘이다. 그러나 이 알고리즘의 탐사 원리가 간편하고 이해가 용이하여 많은 응용 알고리즘이 있다.

2.2.1 Apriori 알고리즘

그림 1은 Apriori 알고리즘이며, 그림 2는 알고리즘 내에서 항목의 조인(join)과 전지(prune)를 수행하는 Apriori-Gen 함수이다. 이 알고리즘의 첫 번째 단계에서는 빈도수를 계산하여 빈발 1-항목집합을 결정하고, k(k≥2)번째는 두 단계로 분할하여 알고리즘이 진행된다.

먼저, (k-1)번째 검색에서는 발견된 빈발항목집합 L_{k-1}를 후보항목집합 C_k으로 만든다. 다음으로 DB를 검색하여 C_k에 있는 후보항목집합의 지지도를 계산한다. 해시트리(hash tree)를 이용하여 지지도 계산의 효율성을 높이고 있다[4]. C_k에 있는 후보항목집합 중에서 최소지지도를 만족하는 항목만 L_k에 진입시킨다. 이러한 시행은 L_k가 더 이상 발견되지 않을 때까지 반복한다. 이 알고리즘의 성능은 그림 2의 조인과 전지에 많은 영향을 받고 있다.

그림 3은 표 1. 예제 데이터베이스를 대상으로 Apriori 알고리즘으로 빈발항목집합을 탐사하는 과정이다. 표 1에는 트랜잭션 4개, 항목 5개의 예제 데이터베이스이며, 최소지지도는 50%로 가정한다. 그러므로 트랜잭션 4개중에서 2개 이상에 해당 항목이 포함되어 있어야 빈발하다고 할 수 있다.

```
// DB 검색하여 C1, L1 생성
L1= {large 1-itemsets}
for (k=2 ; Kk-1 ≠ φ ; k++) do begin
    Ck = apriori-gen(Lk-1) ; // New candidates
    forall transaction t ∈ D do begin
      Ct = subset(Ck, t) ;
        forall candidates c ∈ Ct do
            c.count++ ;
    end
        Lk = { c∈ Ck | c.count ≥ Smin }
end
Answer = ∪k Lk ;
```

그림 1. Apriori 알고리즘

```
1) Join 단계
  insert into Ck
  select p.item1, p.item2, p.item3,··· p.itemk-1,
         q.itemk-1
  from Lk-1 p, Lk-1 q// self join
  where p.item1 = q.item1, ··· p.itemk-2 =
         q.itemk-2, p.itemk-1 < q.itemk-1 ;
2) Prune 단계
  forall itemset c ∈ Ck do
     forall(k-1) - subsets s of c do
        if(s ¬∈ Lk-1)then
          delete c from Ck ;
```

그림 2. Apriori-gen 함수

표 1. 예제 데이터베이스

TID	항목(items)
1	1, 3, 4
2	2, 3, 5
3	1, 2, 3, 5
4	2, 3

위 예를 통하여 알 수 있듯이항목의 존재 유무와 발생 빈도만으로 빈발항목집합을 생성하며, 해당 항목의 수량은 고려하지 않고 있다.

2.3 기존의 수량항목 분할방법

다음과 같은 2개의 연관규칙(R_1, R_2)을 살펴보면, R_1은 나름대로 의미가 있지만, R_2가 R_1보다 고급 정보이다.
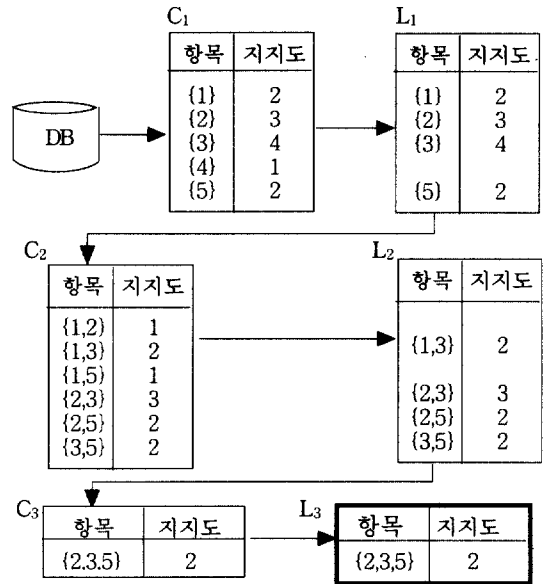
그림 3. Apriori 수행과정

R_1 : "노트와 연필을 구매한 고객들의 85%가 지우개를 동시에 구매한다"

R_2 : "노트 2권과 연필 3자루를 구매한 고객들의 85%가 지우개 1개를 동시에 구매한다"

R_1은 항목의 발생 여부만으로도 연관규칙의 탐사가 가능하지만, R_2는 불가능하다. 그러므로 항목의 값을 임의의 실수(real) 값으로 확장 등의 방법이 필요하게 되어, 최근에 항목들을 범주 데이터(categorical data)과 수량 데이터(quantitative data)로 구분하여 접근하는 연구가 소개되고 있다.

수량 데이터의 연관규칙 탐사 방법으로 수량항목을 이진항목으로 변환하여 기존의 탐사 알고리즘으로 연관규칙 탐사를 고려할 수 있다. 이러한 방법에 대한 기존의 연구[8,9]가 있었다.

Skriant[8]는 수량항목의 정의영역, 즉 도메인을 일정한 범위의 소구간으로 일괄분할(partition)한 후, 이웃한(adjacent) 소구간 분할을 병합(merge)하여 최소지지도를 만족하는 빈발 구간 항목집합을 생성한다.

이 경우에는 수량 항목의 정의 영역에 데이터가 골고루 분포된 경우에는 효과적이지만, 일부 영역에 집중된 경우는 비효율적인 면이 있으므로 이의 해결 방법으로 분포도에 따라 분할하는 유동적 분할법[9]이 발표되었다.

이들 두 가지 방법은 2 단계의 절차를 거쳐 최소지지도를 만족하는 빈발구간 항목집합을 생성한다. 첫째 단계에서는 소간격 분할을 생성하며, 두 번째 단계에서는 최소지지도를 만족할 때까지 이웃 소구간을 합병한다. 분할과 병합에 사용되는 기준은 일정범위 분할법에서는 최소지지도만 사용하고, 유동적 분할법은 최소지지도와 최소분할지지도를 사용하여 분할 및 병합을 실시하였다.

따라서 일정범위 분할법은 데이터의 분포를 고려하지 못하는 점이 약점이고, 유동분할법은 최소분할지지도라는 또 다른 분할기준을 사용함으로써 분할을 위한 부수적인 비교가 필요할 뿐만 아니라사용자가 임의로 최소분할지지도을 설정해야하므로 최적의 기준설정이 어렵다는 문제점이 있다.

3. 빈발구간 항목집합 생성방법

3.1 데이터발생의 지역성

수량 항목은 그림 4와 같이 다양한 유형으로 나타나고 있으며, 이들 항목들의 정의 영역(도메인) 또한 좁은 영역, 넓은 영역 등 다양하다.

(a) 변환이 필요한 수량 항목
- 나 이 :□□□세
- 통근통학 소요시간 :□시 □□분
- 주택의 연건평 :□□□평(㎡) 등

(b) 조사표 상에 분할하여 조사한 수량 항목
- 컴퓨터 활용 상태
① 매일 사용② 1주일 1회 사용 ③ 1달에 1회 사용
④ 1달에 1회 미만 사용⑤ 사용하지 않음
- 현직장 근무연수
① 6개월 미만 ② 6-12개월 미만 ③ 1-3년 미만
④ 3-5년 미만 ⑤ 5-10년 미만 ⑥ 10-15년 미만
⑦ 15-20년 미만⑧ 20년 이상등

그림 4. 센서스 항목중 수량 항목의 예

그림 4의 (b)와 같이 정의 영역의 범위가 좁은 경우는 1 : 1매핑으로 연관규칙을 탐사할 수 있지만, (a)와 같은 경우는 정의 영역이 넓기 때문에적정 간격의 빈발 구간항목으로의 변환이 필요하다. 이때 해당구간의 빈도(frequency)를 근거로 빈발 유무를 판단한다.

그림 5의 (a)는 인구주택총조사의 대전지역 나이별 인구분포, (b)주택의 연건평별 가구분포, (c) 사업체기초통계조사의 종사자규모별 사업체분포를 나타낸다. 이 예를 보면 정의영역 내의 특정 영역을 중심으로 데이터 발생의 빈도가 집중되고 있음을 알 수 있다. 이와 같은 특성은 예의 데이터 외에도 관심 지역, 연령 등에 따라 음반판매, 영화관 관람자 등과 같은 실세계 데이터에서도 나타날 수 있는 특성이다.

이 특성을 데이터 발생의 지역성(locality)라 정의한다. 이의 존재는 빈도가 가장 높은 최빈수(mode)의 단위구간[2]을 기준으로 좌~우로 확장하여 고려대상 구간을 설정하였을 때, 고려대상 구간이 차지하는 비율(A)과 빈도가 차지하는 비율(B)의 비(rate), 즉 B/A의 값으로 판단하며, 지역성의 정도라 정의한다.

그림 5의 데이터 예에서 고려대상 구간을 50%로 하였을 때, (a)에 대한 지역성 정도는 1.42(0.71/0.5), (b)는 1.94(0.97/0.5), (c)는 1.99(0.995/0.5)이다. 이는 최빈수의 단위구간을 중심으로 주변 50%의 구간에 포함된 빈도가 전체 빈도의 71%, 97%, 99.5%에 해당됨을 의미한다. (b)와 (c) 경우는 지역성이 극히 높은 경우로써 빈도 3%, 0.5%가 구간 50%에 달하는 영역을 차지하고 있다. 이러한 구간은 신뢰도를 바탕으로 하는 규칙의 생성 단계에서 낮은 신뢰도로 인하여 효용성이 부족한 구간이다. 따라서 이들 구간에서 생성된 빈발구간 항목은 활용효과가 거의 없다.

이와 같은 데이터 발생 분포의 특성인 지역성을 수량 항목의 연관규칙 탐사에 필요한 빈발구간 항목집합의 생성에 활용하고자 한다.

3.2 지역성을 고려한 빈발구간 항목 생성방법

본 논문에서 제안하는 방법은 최빈수(mode)의 단위구간을 중심으로 수량 항목의 정의영역을 이진항목으로 변환하여 빈발구간 항목집합을 생성하는 방법이다.

3.3.2 빈발구간 항목집합 생성 알고리즘

데이터 발생의 밀집도가 높은 영역이 최빈수의 단위구간이다. 우수한 결과를 얻는 방법으로 이 최빈수의 단위구간을 이용하는 방법이다.

3.3.1 기호설명

빈발구간 항목집합을 생성하기 위하여 필요한 데

---

2) 수량항목 정의영역의 기본단위를 단위구간으로 정의한다.
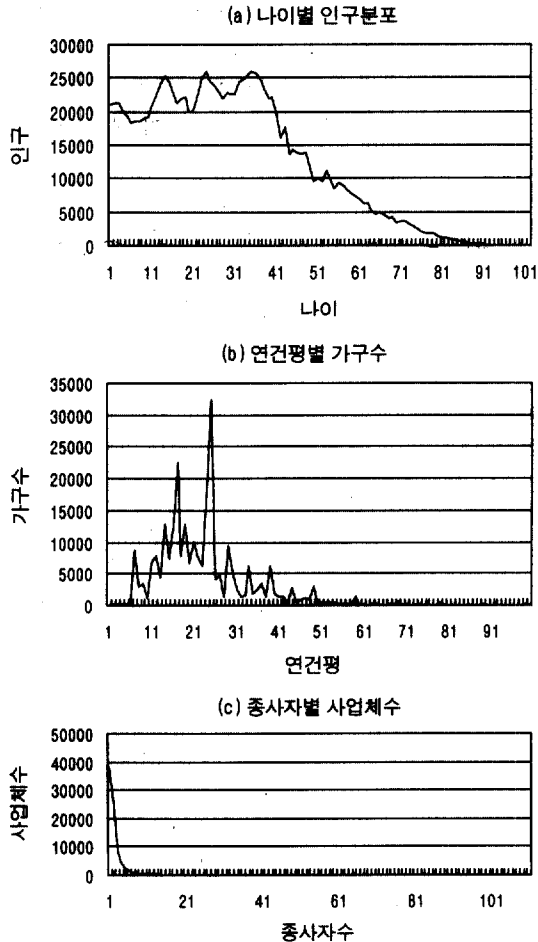
그림 5. 데이터 발생의 지역성(예)

이터베이스(D), 수량 항목(Lq), 발생빈도(f(Lq)) 등의 기호를 다음과 같이 정의한다.

• D는 유한한 범위의 수량 항목이 포함된 트랜잭션들의 집합으로 Lq, f(Lq)를 포함한다.

• Lq는 { $lq_1, lq_2, \dots, lq_{n-1}, lq_n$ }이며, lq_i ($1 \le i \le n$)는 단위구간 항목(item)으로 이산(discrete)하다.

• f(Lq)는 {$f(lq_1), f(lq_2), \dots, f(lq_{n-1}), f(lq_n)$}이며, $f(lq_i)$ ($1 \le i \le n$)는 단위구간에서 데이터 발생빈도이다.

• Max_lq는 최빈수(mode)의 단위구간이다.

• F_L은 빈발구간 항목집합(large interval itemsets)이며, m개의 원소들{$fl_1, fl_2, \dots, fl_m$ }로 구성되고, 각각은 최소지지도(S_{min})를 만족한다.

• lq_t_i($1 \le i \le n$)는 해당 단위구간 lq_i의 사용 가능여부를 표기한다. □

빈발구간 항목집합 생성 방법은 1차 최빈수의 단위구간(lq_i)을 선택한 후, 이를 기준으로 인접(좌~우) 단위구간(lq_{i-1}, lq_{i+1})을 최소지지도를 만족할 때까지 병합한다. 이때 좌~우 항목의 값(빈도 또는 지지도) 중에서 최소지지도 보다 높거나 같으면서 가장 근접하는 값을 취하여 최소지지도를 만족할 때까지 병합한다. 만약 하한한계 또는 상한한계[3]로 인하여 더 이상 진행을 할 수 없을 경우는 한 쪽 값만을 취하여 병, 합을 진행한다. 진행도중에 양쪽(좌~우) 모두 한계에 도달하면 비빈발이므로 빈발하지 않은 구간으로 설정하고, 다음 최빈수의 단위구간을 선정하여 동일하게 진행한다.

인접 단위구간을 병합하는 도중에 최소지지도를 만족하면 병합을 중단하고, 빈발구간 항목집합에 포함시키고, 동시에 빈발구간 항목영역으로 설정한다. 계속하여 잔여 단위구간 중에서 최빈수의 단위구간을 선택하여 동일한 방법으로 빈발구간 항목을 생성하며, 더 이상 단위구간이 존재하지 않을 때 중단한다.

```
// 최소지지도(Smin)는 사용자가 지정
// DB를 검색하여 f(lq)를 생성
 FL = φ
for (k=1 ; Lq≠φ; k++) do begin
    Max_lq = MAX(f(lqi)), (not tagged, 1≤i≤n)
    flk merge lqi ;
    CALL Gen_FL
    FL = U flk// Answer
    Max_lq = 0
    // if sum of lqs between tagged < Smin
       then not large quantitative itemsets
       lqi = used tag (1≤i≤n)// not large
end
```

그림 6. 제안 빈발항목생성 알고리즘

그림 6, 7은 이러한 절차를 코드로 표기한 것이다. 그림 6은 최빈수의 단위구간을 선정하여 Gen-F_L함수에 그 값을 전달한다. 그림 7에서는 전달된 단위구

3) 상한과 하한 한계는 하한(lq_1) 및 상한(lq_n) 경계 또는 이미 사용한 영역의 단위구간 경계이다.

간을 기준으로 최소지지도(S_{min})를 만족할 때까지 좌~우의 단위구간을 병합하는 절차를 수행한다. 이때 상한과 하한 경계의 인접 여부에 따라 4가지 경우(case)를 고려하고 있다.

최빈수의 단위구간을 기준으로 생성된 빈발구간 항목은 그림 8에서 보듯이 1차 최빈수를 기준한 항목의 구간영역이 k차 최빈수를 기준한 항목의 구간영역보다 좁거나 같다. 즉, 1차에서 k차로 진행될수록 빈발구간 항목의 폭이 넓어지는 특징이 있다.

이는 초기에 생성된 빈발구간 항목이 나중에 생성된 빈발구간 항목 보다 데이터 자체가 가지고 있는 특성의 손실이 적은 세밀한 빈발구간 항목임을 의미한다. 그러므로 생성되는 빈발구간 항목의 생성순서는 향후 규칙(rules)을 생성할 때 좋은 정보로 활용될 수 있다.

```
Function Gen_FL
  for (j=1; Max_lq ≥ Smin, j++) do begin
    case 1 : lq_ti-j, and lq_ti+j are not tagged
      if (f(lqi-j)+f(lqi+j)) ≤ (Smin-Max_lq)
        then Max_lq = Max_lq + f(lqi-j) +f(lqi+j);
             flk merge lqi-j, lqi+j ; lq_ti-j, lq_ti+j = tag
        else if (f(lqi-j)≤f(lqi+j) and
                (Smin- Max_lq) ≤ f(lqi-j))
                then Max_lq = Max_lq + f(lqi-j);
                     flk merge lqi-j; lq_ti-j = tag
                else Max_lq = Max_lq + f(lqi+j);
                     flk merge lqi+j; lq_ti+j = tag
           endif
        endif
    case 2 : lq_ti-j is not tagged,lq_ti+j tagged
             Max_lq = Max_lq + f(lqi-j) ;
             flk merge lqi-j;lq_ti-j = tag
    case 3 : lq_ti-j is tagged,lq_ti+j not tagged
             Max_lq = Max_lq + f(lqi+j) ;
             flk merge lqi+j; lq_ti+j = tag
    case 4 : lq_ti-j andlq_ti+jis tagged
             return// flk is not large
  end
Return
```

그림 7. 단위구간 병합함수

빈발구간 항목집합을 생성하는 절차의 예는 표2와 같으며, 10세부터 40세까지의 31개 단위구간에 대하여 데이터 1000건, 최소지지도 10%를 사용하였다.
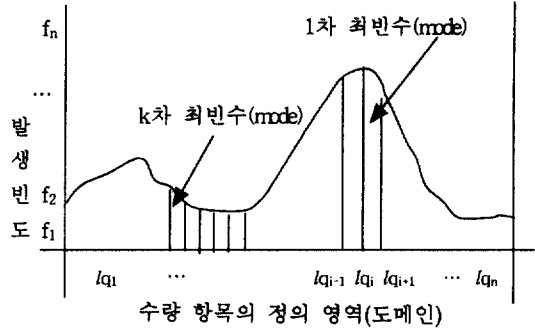
그림 8. 최빈수를 이용하는 방법

그리고 이 데이터 고려대상 구간을 50%로 한 지역성의 정도는 1.12(0.56/0.5)이다.

예의 결과는 빈발구간 항목은 8개, 비빈발 항목은 3개가 생성되었다. 예에서 1차, 2차로 생성된 빈발항목은 2개, 7차는 5개, 8차는 4개의 단위구간이 병합되었다. 그러나 지역성의 정도가 높지 않아 병합된 단위구간의 수는 2개에서 5개로 편차가 그리 크지 않다.

4. 성능평가

본 논문에서 제안한 데이터 발생의 지역성을 고려하여 빈발구간 항목집합을 생성하는 방법의 성능평가 위하여 3가지 방법, 즉 일정범위 분할 및 병합방법(M1), 유동적 분할 및 병합방법(M2) 그리고 제안한 방법(M3)을 사용하여 생성되는 빈발구간 항목 수와 함께 생성구간의 평균간격을 비교한다.

성능평가에는 다음의 3가지 데이터를 사용한다.

i) 인구주택총조사 중 대전광역시의 연령별 인구데이터 1,214,327 레코드

ii) 사업체기초통계조사 중 대전광역시의 종사자 규모별 사업체 데이터 88,869 레코드

iii) 지역성이 없는 모의 데이터 37,000 레코드

이들 데이터의 분포는 그림 5의 (a), (c)와 같으며, 그림 9는 시험데이터 iii)의 분포도로써 지역성이 전혀 없는 데이터이다.

또한 사용한 최소지지도는 9가지(40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, 3%)를 사용하였으며, M1에서 사용한 일정분할 간격은 2, M2의 최소분할지지도는 최소지지도의 1/2로 하였다.

표 2. 빈발구간 항목 생성과정(예)

나이	지지수 (빈도)	단위구간 지지도(%)	최빈수 차수	병합구간 지지도	빈발유무
10	11	1.1		3.4	비빈발
11	23	2.3	10차		
12	25	2.5		11.4	빈발
13	28	2.8			
14	30	3.0			
15	31	3.1	8차		
16	33	3.3		10.8	빈발
17	38	3.8	5차		
18	37	3.7			
19	33	3.3	6차	12.8	빈발
20	32	3.2			
21	32	3.2			
22	31	3.1			
23	32	3.2	7차	11.9	빈발
24	24	2.4			
25	23	2.3			
26	19	1.9			
27	21	2.1			
28	14	1.4	11차	1.4	비빈발
29	21	2.1		10.1	빈발
30	39	3.9			
31	41	4.1	3차		
32	54	5.4		11.3	빈발
33	59	5.9	2차		
34	61	6.1	1차	10.9	빈발
35	48	4.8			
35	39	3.9	4차	11.0	빈발
37	35	3.5			
38	36	3.6			
39	28	2.8	9차	5.0	비빈발
40	22	2.2			

그림 9. 데이터 iii)의 분포도

4.1 생성 빈발구간 항목수 비교

각 방법의 성능평가 결과는 그림 10과 같다. 이들 표를 보면 최소지지도 25%이상에서는 기존 방법(M1, M2)과 제안한 방법(M3)이 동일한 수의 빈발구간 항목을 생성하지만, 25% 미만부터는 제안한 방법이 보다 많은 수의 빈발구간 항목을 생성하고 있음을 알 수 있다.
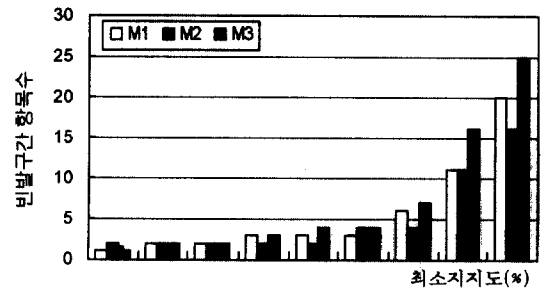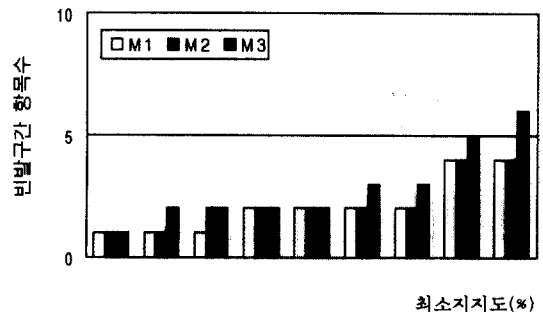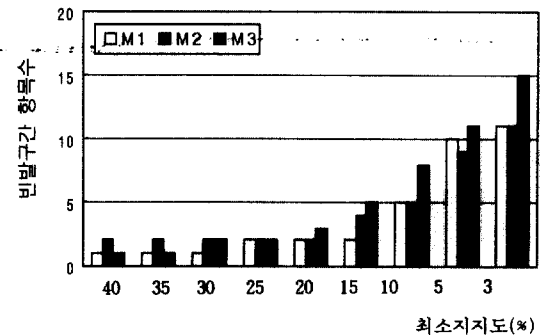
그림 10. 생성 빈발구간 항목수

각 데이터별 시험결과를 살펴보면, 첫째 그림 10 (a)의 경우는 최소지지도 25%에서 기존의 방법과 제안한 방법이 대등하나 최소지지도가 낮아질수록 제안한 방법이 보다 많은 빈발구간 항목을 생성함을 알 수 있다.

둘째로 높은 지역성을 갖는 데이터의 결과인 (b)는 기존의 방법인 M1과 M2가 동일한 반면, M3는 최소지지도 15%부터 빈발구간 항목수가 증가하고 있다.

마지막으로 지역성이 전혀 없는 데이터에 대한 결과인 (c)는 최소지지도 40%, 35%에서는 M2가 우수하고, 20%와 15%에서는 M2와 M3가 비슷하다. 반면에 10%이하에서는 M3가 보다 많은 빈발구간 항목을 생성하고 있다.

이상과 같이 제안한 방법이 기존의 방법보다 최소지지도 25%이상에서는 동등한 성능이지만, 최소지지도 25% 미만에서는 보다 많은 빈발구간 항목을 생성할 수 있으므로 성능의 우수함을 보여준다.

4.2 구간 평균간격 비교

각 방법으로 생성된 빈발구간 항목들의 구간들의 평균간격은 그림 11과 같다. 이들 그래프를 보면 대체로 데이터의 분포 특성에 따라 M1 방법과 M2, M3가 큰 차이가 있음을 알 수 있다. 이는 데이터의 분포를 고려하지 않은 M1방법이 다소 큰 간격의 빈발구간 항목들을 생성하고 있다.

각 데이터별 구간 평균간격을 살펴보면, 첫째 그림 11의 (a)는 최소지지도 25% 미만에서는 큰 차이가 없으나, 25% 이상에서는 데이터 분포의 특성을 고려하지 않은 M1의 구간 평균간격이 타 방법에 비하여 월등히 넓음을 알 수 있다. 그리고 데이터 분포의 특성을 고려한 방법인 M2와 M3는 최소지지도에 따라 차이는 있지만 대체로 M3가 좁은 구간 평균간격의 빈발구간항목집합을 생성함을 알 수 있다.

둘째, 표 11의 (b)는 높은 지역성의 데이터에 대한 빈발구간 항목들의 구간 평균간격 비교 그래프이다. 데이터의 분포특성으로 인하여 M1, M3의 평균구간간격이 특이하게 나타나고 있다. M1은 최소지지도 30%에서 15%사이에 넓은 구간 평균간격이 보이고 있다. 이는 데이터 분포의 특성을 고려하지 않았기 때문이다. M3의 경우 '*특이 부분*'이라 표시한 부분에 유달리 넓은 구간 평균간격을 보이고 있다. 이는 앞(3.1절)에서 언급한 바와 같이 빈도가 극히 낮은 영역에서 생성된 빈발구간 항목으로 인하여 구간 평균간격이 넓게 나타나고 있다. 이러한 경우는 생성순서 정보를 활용하여 제거하거나 규칙생성 과정에서 활용유무 판단이 필요하다.

M3'은 M3가 생성한 빈발구간 항목들 중에서 나중에 생성된 것을 제외하여 M2와 동일한 수의 빈발구간 항목에 대한 구간 평균간격이다. 그 결과 M3'는 M2와 거의 동일한 구간 평균간격을 보이고 있다.
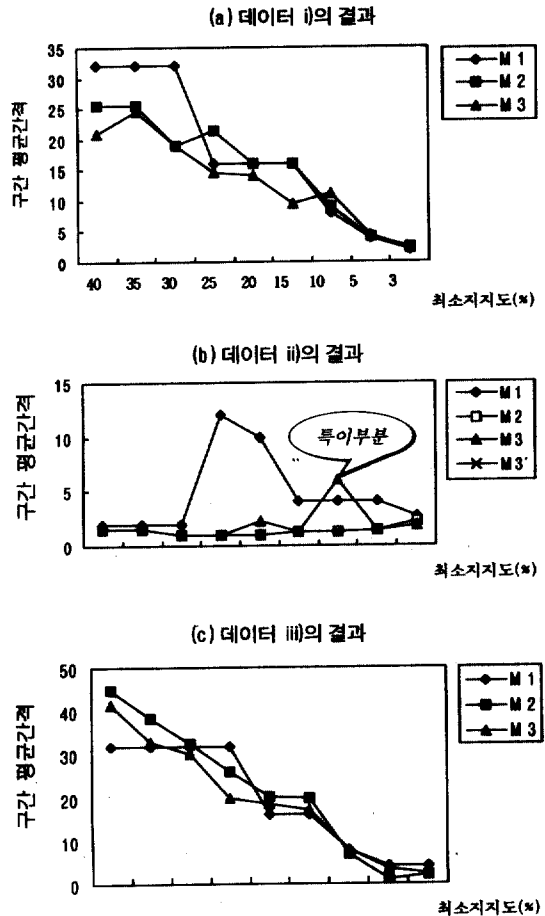
그림 11. 구간 평균간격

마지막으로 그림 11의 (c)는 지역성이 없는 모의 데이터에 대한 구간 평균간격의 그래프이다. 이 데이터에 대해서는 세 가지 방법 모두의 구간 평균간격이 유사하게 나타나고 있다. 즉 최소지지도에 따라 구간 평균간격이 넓어지기도 하고 좁아지기도 한다. 이러한 데이터 분포는 많은 빈발구간 항목을 생성하는 방법이 우수한 방법이다.

4.3. 성능평가 종합

이상과 같이 각 방법에 대하여 생성되는 빈발구간

항목수와 생성된 빈발구간 항목들의 구간 평균간격을 비교하여 보았다.

이들을 종합하여 보면, 생성되는 빈발구간 항목수는 최소지지도 25%를 기준으로 그 이상에서는 세 가지 방법이 대등하지만, 구간 평균간격은 M1의 월등히 넓게 나타나고 있고, 다음에 M2, M3 순으로 나타나고 있어 제안한 방법이 우수함을 알 수 있다. 반면에 최소지지도 25% 미만에서는 제안한 방법이 많은 수의 빈발구간 항목을 생성하고 있지만, 데이터의 분포 특성으로 인하여 구간평균간격이 넓어지는 경우가 발생하였다. 이는 많은 수의 빈발구간 항목을 생성하여도 모든 항목이 세밀하지 않을 수 있음을 의미한다. 그러나제안한 방법은 생성되는 순서 정보도 동시에 생성하기 때문에 이를 이용하여 효용성이 떨어지는 구간을 제고할 수 있다. 즉 초기에 생성된 빈발구간 항목은 세밀한 항목이며, 그 이후 생성 순서에 따라 항목의 세밀도가 조금씩 떨어지는데, 사용자는 이들 순서정보를 적절히 활용함으로써 보다 우수한 연관규칙을 발견할 수 있기 때문이다.

5. 결론 및 향후과제

본 논문에서는 수량 항목을 포함하는 대용량의 데이터베이스에서 연관규칙의 탐사를 위해 수량 항목의 정의영역을 이진항목 형태의 빈발구간 항목으로 변환하는 보다 효율적인 방법을 제안하였다. 그리고 인구주택총조사 등 실제 데이터를 사용하여 성능평가를 실시하여 제안한 방법이 기존의 방법보다 보다 많은 수의 세밀한 빈발구간 항목집합을 생성할 수 있으므로 성능의 우수함을 알았다.

제안한 방법은 탐사 대상 데이터가 가지는 특성, 즉 데이터 발생의 지역성을 고려하는 방법으로 최빈수를 빈발구간 항목생성을 위하여 사용하였다. 최빈수를 사용함으로써 얻는 효과로는 보다 세밀한 빈발구간 항목을 생성함은 물론 그림8에서와 같이 최빈수의 차수, 즉 생성순서에 따라 빈발구간 항목의 세밀도가 감소하는 특징이 있다. 이는 생성되는 빈발구간 항목의 순서에 따라 원 데이터가 가지고 있는 특성의 손실 정도가 다르다는 것이다. 즉 초기에 생성된 것이 나중에 생성된 것 보다 손실이 적은 특징을 가진다. 이 특징은 향후 연관규칙을 탐사할 때, 필요한 규칙의 질(quality)의 정도에 따라 사용자가 빈발구간 항목의 생성순서를 감안하여 활용유무를 조정할 수 있다. 다만, 이 때 고려하여야 할 사항은 최빈수가 지역성이 없는 영역에 존재하는 특이한 분포의 데이터일 경우는 초기에 생성된 빈발구간 항목이 앞서 언급한의 특징을 갖지 못할 수 있다는 것을 감안하여야 한다.

향후의 연구 과제로는 첫째, 앞에서 언급한 바와 같이 빈발구간 항목의 생성순서를 연관규칙의 탐사에 효과적으로 활용하는 방법에 대한 연구이다. 둘째 과제는 3.1절의 예에서 본 바와 같이 조사 당시에 분할된 수량항목의 경우에 대한 빈발구간 항목집합의 생성방법에 대한 연구도 필요하다.

참 고 문 헌

[1] Sholom M. Weiss, Nitin Indurkhya, *Predictive Data Mining*, Morgan Kaufmann Publishers, Inc., San Francisco, California, 1998.

[2] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", *In Proc. of the ACM SIGMOD Conference on Management Data*, pp. 207-216, 1993.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for mining association rules", *In Proceedings of the 20th VLDB Conference*, Santiago, Chile, Sept., 1994

[4] J.S. park, M.S. Chen and P.S. Yu, "An Effective hase-based algorithm for mining association rules", *In Proceedings of ACM SIGMOD Conference on Management of Data*, pp. 175-186, May 1995.

[5] A. Savasere, E. Omiencinsky and S. Navathe, "An efficient algorithm for mining rules in large databases", *In proceedings of the 21st VLDB Conference*, pp. 432-444, 1995.

[6] J.S. Park, P.S. Yu and M. S. Chen, "Mining Association Rules with Adjustable Accuracy", *In Proceedings of ACM CIKM 97*, pp. 151-160, November 1997.

[7] M.J. Zaki, S. Parthasarathy, Wei Li, and M. Ogihara, "Evaluation of Sampling for Data Min-

ing of Association Rules", The Univ. Rochester Technical Report 617, May 1996.

[8] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of the ACM SIGMOD Conference on Management of Data*, June 1996.

[9] 최영희, 장수민, 유재수, 오재철, "수량적 연관규칙탐사를 위한 효율적인 고빈도항목열 생성기법", 한국정보처리학회 논문지 제6권 제10호, pp. 2597-2607, 1999.

[10] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo, "Fast Discovery of Association Rules", *In Advances in Knowledge Discovery and Data Mining*, AAAI Press, pp. 307-328, 1996

[11] S. Brin, R. Motwani, J.D. Ullman and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", *In Proceedings of ACM SIGMOD Conference on Management of Data*, pp. 255-264, May 1997.

박 원 환

1982년 충남대학교 전산학과(이학사)
1986년 충남대학교 대학원 전산학과(이학석사)
1996년 전자계산조직응용 기술사
1997년~ 현재 순천향대학교 대학원 박사과정
1988년~현재 통계청 전산서기관
관심분야 : 병렬처리, 데이터마이닝, 데이터베이스

박 두 순

1981년 고려대학교 수학과(이학사)
1983년 충남대학교 대학원 전산학과(이학석사)
1996년 고려대학교 전산학 전공(이학박사)
1992년~1993년 미국 U. of Illinois at Urbana-Chamaign CSRD 객원교수
2000년~현재 순천향대학교 컴퓨터교육원 원장
1985년~현재 순천향대학교 정보기술공학부 교수
관심분야 : 병렬처리, 데이터마이닝, 정보검색, 멀티미디어